Sparks.

Programme de formation
Data Science Fondamentaux

• Objectifs

Cette formation constitue une initiation aux concepts, principes et outils de la science des données. Vous explorerez le processus de Data Science de bout en bout, et découvrirez les principales techniques mises en oeuvre au quotidien par les data scientists. Préparation des données, visualisation, modélisation, analyse, restitution... à l'issue de notre formation vous maîtriserez l'ensemble des bonnes pratiques pour manipuler des ensembles de données et saurez les mettre à profit dans un contexte professionnel. La formation alterne entre apports théoriques, mises en pratiques et échanges sur les différents contextes des stagiaires afin de faciliter l'acquisition de savoirs. Travaillez sur des problèmes concrets de Data Science, avec des données réelles, et favorisez une prise de décision éclairée au sein de votre organisation grâce à notre formation Data Science de référence !

• Pré requis

Connaissances en programmation, statistiques et probabilités.

• Durée

5 jours

• Public

Analystes, Architectes, Chefs-de-projet, Developpeurs, Managers

•Plan de formation

Introduction à la formation Data Science Fondamentaux

L'avènement de la data, nouvelle ressource stratégique pour les entreprises
Qu'entend-on par Big Data ? Architectures, stockage, traitement...
La règle des 3V : Volume, Vélocité et Variété
Cas d'usage et domaines d'application des solutions Big Data
De l'analyse statistique au deep learning : retour historique sur le traitement des données
Data Mining vs. Business Intelligence
Enjeux, perspectives et défis pour les entreprises, organisations et Etats
Gouvernance des données : cycle de vie et gestion de la qualité

Principes et concepts de base en Data Science

Qu'est-ce que la Data Science ? Introduction à la science des données
Définitions, terminologie : le vocabulaire de la Data Science
Data Scientist, « métier le plus sexy du XXIème siècle » ?
Comprendre le rôle, les compétences et la pensée du data scientist
Vue d'ensemble d'un processus de Data Science
Comprendre ce qu'est le Data Mining
Identifier le besoin et les objectifs métiers

La boîte à outils du Data Scientist

Panorama des outils open-source et propriétaires du marché
Les langages R, Python et leur environnement de développement (RStudio IDE, Anaconda...)
Travailler avec les notebooks Jupyter
Les principales bibliothèques pour la Data Science : Pandas, NumPy, SciKit-Learn...
Bases de données : SQL, NoSQL, MongoDB...
Visualisation : Excel, Tableau, Matplotlib, D3.js...
Installer les outils nécessaires aux travaux pratiques de la formation

Rappels mathématiques, statistiques et probabilités

Programmation avec R ou Python
Présentation d'un langage de programmation pour la Data Science
Caractéristiques du langage, structure d'un programme
Assigner des variables, types de données, opérations de base
Manipuler des listes, tableaux, fonctions, packages...

Obtention et exploration des données
Où trouver des ensembles de données ?
Sources de données publiques et privées (web, médias sociaux, IoT...)
Les entrepôts de données (datawarehouse, datalake)
Importer des données, installer des packages et des bibliothèques
Une première visualisation : identifier les caractéristiques d'un ensemble de données
Quelles sont les données pertinentes ?
Données opérationnelles
Bonnes pratiques pour contrôler la qualité des données

Prétraitement de données
Comprendre l'importance du processus de nettoyage des données
Exemple d'un ensemble de données non-structurées
Nettoyer et préparer des ensembles de données
Identifier et gérer les valeurs manquantes ou aberrantes
Considérations pour le Big Data : les outils Apache Spark, Hadoop et le modèle MapReduce
L'analyse en composantes principales (ACP, ou PCA pour Principal Component Analysis)
Feature engineering : extraction et sélection des features

Analyse et modélisation : introduction au Machine Learning
Modéliser un problème de Data Science : entrées et sorties attendues
Le Machine Learning et les capacités d'apprentissage des machines
Les différentes familles d'algorithmes : supervisé, non-supervisé, semi-supervisé, classification, régression...
L'intuition derrière un modèle d'apprentissage
Bibliothèques et packages ML pour R et Python : scikit-learn, gradDescent, TensorFlow...
Analyse et exploration statistiques de documents : le Text Mining
Gérer les gros volumes de données (Big Data)

Mise en oeuvre des méthodes d'apprentissage supervisé
Estimation de valeurs : construire un modèle de régression linéaire
Régression non-linéaire, régression logistique
Interpréter les coefficients de régression
Utiliser l'algorithme du gradient (descente de gradient)
Automatiser la labélisation de nouveaux jeux de données
Vue d'ensemble des méthodes ensemblistes
Réseaux Bayésiens, classification naïve bayésienne
Arbres de décision et random forests
Machines à vecteurs de support (SVM)

Apprentissage semi-supervisé et non-supervisé, clustering
Les principaux algorithmes
Partitionnement en k-moyennes
Regroupement hiérarchique
Clustering basé sur la densité
Qu'est-ce que le Deep Learning ? Présentation des réseaux de neurones

Evaluation et tests des modèles d'apprentissage
Evaluer et améliorer des modèles : sur-apprentissage, cross-validation...
Métriques et méthodes pour la maintenance des modèles
Pourquoi la performance des modèles d'apprentissage se détériore-t-elle ?
Ajuster et valider un modèle

Visualisation et restitution : communiquer avec les données
Transformer des données en décisions
Les principes de la visualisation de données
Outils principaux de dataviz : Tableau Software, QlikSense...
Représentations graphiques de base :

histogrammes, boxplots et diagrammes
Les packages R pour la datavisualization (R Markdown, Shiny...)
Visualisation interactive de données
Data storytelling : raconter une histoire avec les données

Défis et opportunités : la Data Science dans votre organisation
Intégrer la Data Science dans les processus actuels
Sélectionner les bons outils suivant les objectifs et le contexte professionnel
Enjeux organisationnels, éthiques et juridiques